Statistics for Biology and Health

Borchers/Buckland/Zucchini: Estimating Animal Abundance: Closed Populations.
Burzykowski/Molenberghs/Buyse: The Evaluation of Surrogate Endpoints.
Everitt/Rabe-Hesketh: Analyzing Medical Data Using S-PLUS.
Ewens/Grant: Statistical Methods in Bioinformatics: An Introduction. 2nd ed.
Hougaard: Analysis of Multivariate Survival Data.
Keyfitz/Caswell: Applied Mathematical Demography, 3rd ed.
Klein/Moeschberger: Survival Analysis: Techniques for Censored and Truncated Data, 2nd ed.
Kleinbaum/Klein: Survival Analysis: A Self-Learning Text.
Kleinbaum/Klein: Survival Analysis: A Self-Learning Text. 2nd ed.
Kleinbaum/Klein: Logistic Regression: A Self-Learning Text, 2nd ed.
Lange: Mathematical and Statistical Methods for Genetic Analysis, 2nd ed.
Manton/Singer/Suzman: Forecasting the Health of Elderly Populations.
Moyé: Multiple Analyses in Clinical Trials: Fundamentals for Investigators.
Parmigiani/Garrett/Irizarry/Zeger: The Analysis of Gene Expression Data: Methods and Software.
Salsburg: The Use of Restricted Significance Tests in Clinical Trials.
Simon/Korn/McShane/Radmacher/Wright/Zhao: Design and Analysis of DNA Microarray Investigations.
Sorensen/Gianola: Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics.
Stallard/Manton/Cohen: Forecasting Product Liability Claims: Epidemiology and Modeling in the Manville Asbestos Case.
Therneau/Grambsch: Modeling Survival Data: Extending the Cox Model.
Vittinghoff/Glidden/Shiboski/McCulloch: Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models
Zhang/Singer: Recursive Partitioning in the Health Sciences.

Tomasz Burzykowski
Geert Molenberghs
Marc Buyse
Editors

The Evaluation of Surrogate Endpoints

With 57 Illustrations

Springer

9

Extensions of the Meta-analytic Approach to Surrogate Endpoints

Mitch Gail

9.1 Introduction

Whether an endpoint S is a good surrogate for a true clinical endpoint T depends on the intended use of the surrogate. Our primary goal is to use a surrogate in a clinical trial to estimate the trial-level effect of a new treatment on T without having to measure T. Another possible use of a surrogate is to predict the outcome T on an individual patient.

For clinical management of an individual patient, it would be valuable if S could be used to predict that individual's outcome T reliably, regardless of what treatment, Z, or other covariates, X, might be present. This assumption that T be conditionally independent of Z (and X) given S is the essential component in Prentice's (1989) criteria that define a good surrogate for hypothesis testing. This assumption holds if S is on the sole causal pathway leading to T, and all factors that influence T do so only through their effects on S. Although this strong assumption and ancillary conditions guarantee the validity of hypothesis tests for no treatment effect, they do not insure that S can predict T well at the individual level. Instead, Buyse, Molenberghs, Burzykowski, Renard, and Geys (2000a), which we abbreviate BMBRG, propose the within individual squared correlation, R^2_{indiv}, of T on S as a measure of the adequacy of S for predicting an individual's outcome (see also Chapter 7).

If S could be shown to satisfy the conditional independence assumption and to have a high R^2_{indiv}, one would have powerful evidence for a causal biological role for S and its close biological connection to T. Moreover, one could hope not only to test for treatment effects on T based on those on S, but also to estimate treatment effects on T from those on S. For example, suppose one wishes to estimate $\delta = E(T|Z = 1) - E(T|Z = 2)$ where

$Z = 1$ corresponds to an experimental treatment and $Z = 2$ to a control or standard treatment, possibly a placebo. We assume $Z = 1$ or 2 is assigned at random with equal probability. Suppose a previous study on control subjects has been done that yields an estimate of the density $f(T|S, Z = 1, X)$ that equals $f(T|S)$ by the conditional independence assumption. In the new study population

$$\delta = \int tf(t|s)h(s|Z = 1, x)dG(x)dt - \int tf(t|s)h(s|Z = 2, x)dG(x)dt,$$

where $h(s|z, x)$ is the conditional density of S given Z and X, and $G(x)$ is the distribution function of X. Because $f(t|s)$ is assumed known from previous studies on (T, S) and because $h(s|Z = 2, x)$ and $h(s|Z = 1, x)$ are estimable from the current study using the surrogate endpoint only, one can calculate the effect of the treatment Z on T in this new study without measuring the true clinical endpoint T.

All this depends on the strong conditional independence assumption $T \amalg Z$, X given S, however. It is impossible to verify this assumption empirically, because one would need to examine an infinite number of treatments and covariates. Even for a single study and treatment comparison, there is limited ability to rule out a dependence of T on Z given S with regression methods, leading Freedman, Graubard, and Schatzkin (1992) and Lin *et al.* (1997) to explore the related criterion of percentage of the treatment effect explained (see Chapter 5 for a discussion of this criterion and allied concepts). But without conditional independence, some other basis is needed to attain the central goal of estimating the magnitude of the treatment effect on T in a new trial from data on S only.

The meta-analytic approach to evaluating surrogate markers, introduced by Daniels and Hughes (1997) and BMBRG, leads to an empirical assessment of how well a surrogate can be used to estimate trial-level treatment effects on T. The basic idea is that one can use information from previous similar studies in which both T and S are measured in treated $(Z = 1)$ and control $(Z = 2)$ groups to learn how well the treatment effect on T is predicted by outcomes S in the treated and control groups. In a trial of a new treatment similar to those in the previous studies, one measures only the effects of Z on S and uses data from the previous studies and from the results on S in the new study to estimate the effects of Z on T.

In order to carry out this program, one needs to posit a superpopulation of similar trials from which the new trial and the previous trials are drawn. For example, Daniels and Hughes (1999) studied various retroviral therapies against HIV/AIDS. In some applications it may be unclear whether the new trial with its new experimental treatment is similar enough to previous studies and their treatments to regard it as a sample from the same superpopulation of trials. Even if there is agreement on the class of similar trials, a serious practical limitation may be the small number of previous trials with data on T and S. One relies on superpopulation parameters, which reflect trial to trial variation, in order to infer trial-level treatment effects on T from those on S. Having too few previous trials limits the precision with which superpopulation parameters can be estimated and hence the precision of meta-analytic inference (Gail, Pfeiffer, van Houwelingen, and Carroll 2000, which we abbreviate GPHC).

A second meta-analytic issue concerns the degree to which models describe the joint distribution of T and S at the individual level. Chapters 7 and 10–14 in this book present such detailed models. GPHC describe a marginal approach in which the distributions of S given Z, and T given Z, are modeled separately. They argue that this approach allows great flexibility for describing trial-level treatment effects and avoids having to specify the joint distribution of T and S given Z, which may be poorly understood. The marginal approach also captures most of the available information about trial-level treatment effects. Tibaldi *et al.* (2003) show that estimates of the proportion of variability in the estimated trial-level treatment effect that is explained by the surrogate, R^2_{trial}, is almost identical for marginal ("univariate") and bivariate linear models, as discussed further in Section 9.3.

In Section 9.2 we illustrate these concepts for normal models for S and T, in Section 9.3 we discuss the flexibility of the marginal model approach, and in Section 9.4 we recount some potential practical and theoretical limitations of the meta-analytic approach.

9.2 The Normal Model

Many of the previous ideas are illustrated by the normal model. Let T_{zij} denote the true clinical response of patient j $(j = 1, 2, \ldots)$ in trial i on treatment $Z = z$ $(z = 1$ or $2)$ and define S_{zij} similarly for the surrogate. Here j ranges from 1 to n_i for $Z = 1$ and from 1 to m_i for $Z = 2$. Given $\theta_i = (\theta_{1T_j}, \theta_{1S_j}, \theta_{2T_j}, \theta_{2S_j})^T$, the vector $(T_{1ij}, S_{1ij}, T_{2ij}, S_{2ij})^T$, is normally distributed with mean θ_i and variance-covariance matrix Σ_i, which is block diagonal with non-zero components Σ_{11i} and Σ_{22i}, corresponding respectively to $(T_{1ij}, S_{1ij})^T$ and $(T_{2ij}, S_{2ij})^T$, which are independent. The θ_i come from a normal superpopulation with mean μ and variance ϕ. This model is very similar to that of BMBRG except that it allows for $\Sigma_{11i} \neq \Sigma_{22i}$, whereas BMBRG require $\Sigma_{11i} = \Sigma_{22i}$.

A series of N "previous" trials permits one to estimate the parameters of the superpopulation, μ and ϕ. Within the ith such trial, the mean is estimated as $\widehat{\theta}_i = (T_{1i}, S_{1i}, T_{2i}, S_{2i})^T$, where, for example, $T_{1i} = n_i^{-1} \sum_j T_{1ij}$.

The quantities Σ_{11i} and Σ_{22i} are estimated from the within-trial empirical variance-covariance matrices of $(T_{1ij}, S_{1ij})^T$ and $(T_{2ij}, S_{2ij})^T$, respectively. Because θ_i is normally distributed with mean μ and variance-covariance matrix $\phi + \Sigma_i$, various methods such a maximum likelihood, REML or empirical Bayes can be used to estimate μ and ϕ.

Now suppose we consider a new trial $(i = 0)$ drawn from the superpopulation and only get to observe (S_{10j}, S_{20j}), which have within trial components of variance σ_{220} from Σ_{11i} and σ_{440} from Σ_{22i}. We seek to estimate θ_0 and especially the components that correspond to the unmeasured clinical outcomes T. Let $\theta_{T0} = (\theta_{1T0}, \theta_{2T0})^T$ be the means of T_{10j} and T_{20j}, respectively, and let $\theta_{S0} = (\theta_{1S0}, \theta_{2S0})^T$ be the means of S_{10j} and S_{20j}, respectively. Because $(\theta_{T0}^T, \widehat{\theta}_{S0}^T)^T$ is multivariate normal, the conditional mean and variance of θ_{T0} can be expressed in terms of $\widehat{\theta}_{S0}$ and parameters $\psi = (\mu, \phi, \sigma_{220}, \sigma_{440})$. Indeed, letting D and W be known matrices defined so that $\theta_{T0} = D\theta_0$ and $\theta_{S0} = W\theta_0$ (see Section 2 of GPHC for details),

$$E(\theta_{T0} \mid \widehat{\theta}_{S0}) = D\mu + D\phi W^T[W(\phi + \Sigma_0)W^T]^{-1}(\widehat{\theta}_{S0} - W\mu) \tag{9.1}$$

and

$$\operatorname{Cov}(\theta_{T0} \mid \widehat{\theta}_{S0}) = D\phi D^T - D\phi D^T[W(\phi + \Sigma_0)W^T]^{-1}W\phi D^T, \tag{9.2}$$

where (9.1) and (9.2) only depend on the elements σ_{220} and σ_{440} of Σ_0. The variances σ_{220} and σ_{440} can be estimated from the empirical variances of S_{10j} and S_{20j}, respectively, and μ and ϕ can be estimated from the previous trials. Assuming the elements of ψ are known, one knows the distribution of the means of the unmeasured true clinical outcomes θ_{T0} from the conditional normal distribution defined by (9.1) and (9.2). In particular, for $R = (1, -1)$, one can calculate the distribution of the treatment effect $\delta_0 \equiv R\theta_0 \equiv \theta_{1T0} - \theta_{2T0}$, which is normal with mean $M(\psi) \equiv RE(\theta_{T0} \mid \widehat{\theta}_{S0})$ and variance $V(\psi) \equiv R\operatorname{cov}(\theta_{T0} \mid \widehat{\theta}_{S0})R^T$, which can be calculated easily from (9.1) and (9.2).

If no measurements on the surrogate were available in the new study, but if the parameters of the superpopulation were known without error from many similar previous studies, one could still estimate the new treatment effect as $\mu_{1T} - \mu_{2T}$, with variance $RD\phi D^T R^T$. The proportion by which this variance is reduced by measuring the surrogate in the new study is, from equation (9.2),

$$R^2_{\text{trial}} = \frac{RD\phi W^T[W(\phi + \Sigma_0)W^T]^{-1}W\phi D^T R^T}{RD\phi D^T R^T}. \tag{9.3}$$

If σ_{220} and σ_{440} are negligible, so that Σ_0 is omitted from (9.3), this definition of R^2_{trial} reduces to that given by BMBRG. BMBRG propounded the version of R^2_{trial} (with $\Sigma_0 = 0$) as a measure of the adequacy of the surrogate S at the trial level.

The difference $\delta_0 = \theta_{1T0} - \theta_{2T0}$ is a natural measure of treatment effect, but the distribution of an arbitrary treatment effect function $\delta_0 = \delta(\theta_{1T0}, \theta_{2T0})$ can be obtained analytically or by simulating from the conditional normal distribution of θ_{T0} given ψ and $\widehat{\theta}_{S0}$. An estimate of δ_0 might be $\widehat{\delta}_0 = \delta[E(\theta_{1T0}|\psi, \widehat{\theta}_{S0}), E(\theta_{2T0}|\psi, \widehat{\theta}_{S0})]$, and confidence intervals could be based on the quantiles of the distribution of δ_0 given ψ and $\widehat{\theta}_{S0}$.

9.2.1 *Precision of Estimates of $\boldsymbol{\delta_0}$ Based on the Meta-analytic Approach*

Using the surrogate to estimate the true treatment effect δ_0 can lead to severe loss of precision compared to measuring T directly. Even if a large number of previous trials have been conducted so that μ and ϕ are known without error, and even if the sample size in the new trial on the surrogate tends to infinity, so that $\sigma_{220} = \sigma_{440} = 0$, there is irreducible variability in $\widehat{\theta}_0$ that reflects trial-to-trial variation in θ_i in the superpopulation, as quantified by ϕ. For example, with $\delta_0 = R\theta_{T0}$ defined as above, the variance of $\widehat{\theta}_0$ is

$$RD\phi D^T R^T - RD\phi W^T(W\phi W^T)^{-1}W\phi D^T R^T,$$

which is strictly positive unless θ_{1Ti} and θ_{2Ti} are linearly dependent on θ_{1Si} and θ_{2Si}. In contrast, measuring true endpoints T will yield an estimate of δ_0 with variance tending to zero.

A realistic assessment of the variability of $\widehat{\theta}_0$ also needs to acknowledge uncertainty in $\widehat{\mu}$ and $\widehat{\phi}$, the estimates of superpopulation parameters. GPHC considered a 95% confidence interval on $\delta_0 = \theta_{1T0} - \theta_{2T0}$. A naïve 95% confidence interval that assumes known $\psi = (\mu, \phi, \sigma_{220}, \sigma_{440})$ is $M(\psi) \pm 1.96V^{1/2}(\psi)$ with M and V as defined previously. For $N = 5, 10, 25, 50$ and 100 previous trials, this naïve confidence interval had coverage 0.64, 0.61, 0.82, 0.90 and 0.92 respectively. Thus, with a small number of previous trials, confidence intervals that assume ψ is known without error have subnominal size and can be seriously misleading. GPHC provide bootstrap procedures that give confidence intervals with nominal coverage. These intervals ranged from 4% to 293% longer than the naïve confidence interval, however, as the number of previous trials decreased from $N = 100$ to $N = 5$.

To illustrate further the loss in precision from the meta-analytic approach, GPHC discussed a comparison of pravastatin $(Z = 1)$ with placebo $(Z = 2)$ on a true clinical outcome (T), namely change in coronary artery diameter over a two-year period, and on a surrogate (S), change in total cholesterol. The example was favorable to the meta-analytic approach because, rather than take different trials of similar agents ("statins") from the literature, GPHC chose 10 centers from a single trial, the REGRESS Trial (Jukema *et al.* 1995) as the "previous" studies, and one remaining center as the "new" study. Because all centers were using the same protocol and studying the exact same agent, there was probably less "between-trial" variability, captured in ϕ, than would be expected in a real meta-analysis based on different trials with different agents. Using the clinical endpoint T, the "new study" indicated a favorable treatment effect on decreases in coronary diameter of $\widehat{\theta}_{1T0} - \widehat{\theta}_{2T0} = 0.0381$ mm with 95% confidence interval $[-0.0138, 0.0900]$. Based on the surrogate data only in the "new study", GPHC estimated the true treatment effect as 0.0402 with naïve confidence interval $[-0.0552, 0.1355]$ and with bootstrap confidence interval that takes variation of ψ into account: $[-0.1346, 0.2149]$. Thus, there is a huge loss in precision from relying on S to estimate treatment effects on T.

9.3 Flexibility of the Marginal Approach

In Section 9.2, we made no mention of the ability of the surrogate to predict individual outcomes, which can be assessed in each trial by examining correlations between T and S in Σ_{11i} and Σ_{22i}. The quantities Σ_{11i} and Σ_{22i}, however, only influence estimates of trial-level treatment effects through their impact on estimating μ and ϕ in the superpopulation model and through σ_{220} and σ_{440}. Especially if all the component trials are large, Σ_{11i}, Σ_{22i}, σ_{220}, and σ_{440} have little influence on superpopulation parameters, and inference on trial-level effects is unrelated to how well S predicts T at the individual level. Because the main interest is in estimating effects on T at the trial level, and in order to avoid specification of the joint distribution of T and S, GPHC adopted a marginal approach to modeling.

Suppose θ_{zTi} represents some feature(s) of the marginal distribution of T in treatment group z in trial i, such as the mean, and define θ_{zSi} similarly for features of the marginal distribution of S. Assume that the components of $\theta_i = (\theta_{1T_j}, \theta_{1Sj}, \theta_{2Tj}, \theta_{2Sj})^T$ satisfy separate estimating equations

$$\sum_{j=1}^{n_i} U_{1Tij}(\theta_{1Ti}) = 0, \qquad \sum_{j=1}^{n_i} U_{1Sij}(\theta_{1si}) = 0,$$

$$\sum_{j=1}^{m_i} U_{2Tij}(\theta_{2Ti}) = 0, \qquad \sum_{j=1}^{m_i} U_{2Sij}(\theta_{2Si}) = 0.$$

We assume that U_{1Tij} is functionally independent of θ_{1Si}, θ_{2Ti}, and θ_{2si}, and that other estimating equations likewise depend only on the parameters shown in their arguments. As in GPHC, it is possible to estimate within experiment variance-covariance matrices Σ_i, namely the conditional covariance of $\widehat{\theta}_i$ given θ_i, from the empirical covariances of terms like $U_{1T_{ij}}$ and U_{1Sij}. Moreover, if θ_i is drawn from a normal $N(\mu, \phi)$ superpopulation, the methods in Section 9.2 can be applied to obtain inference on $\delta_0 = \delta(\theta_{1T0}, \theta_{2T0})$.

The marginal approach is very flexible. For example, if T and S are dichotomous with values 1 or 0, we might choose θ_{zSi} to be the logarithms of the marginal odds that $T = 1$ on treatment z in trial i and θ_{zSi} to be marginal odds that $S = 1$. Inference on the log odds ratio, $\delta_0 = \theta_{1T0} - \theta_{2T0}$ follows directly from (9.1) and (9.2) with allowance for uncertainty in ψ. The risk difference

$$\delta_0 = \exp(\delta_{1T0})/[1 + \exp(\delta_{1T0})] - \exp(\theta_{2T0})/[1 + \exp(\theta_{2T0})]$$

is non-linear in θ_{1T0} and θ_{2T0}, and inference can be based on simulations from the conditional distribution of θ_{T0} given $\widehat{\theta}_{S0}$, with allowance for uncertainty in ψ, as in GPHC.

Marginal models can also be used for survival data. For example, T_{zij} might have a Weibull distribution, $P(T_{zij} \leq y) = 1 - \exp(-\lambda_{2Ti} y^{\alpha_{zTi}})$. Likewise, S_{zij} might have a Weibull distribution with parameters λ_{zSi} and α_{zSi}. The alternative parameters $\theta_{zTi} = (\ln(\lambda_{zTi}), \alpha_{zTi})^T$ and $\theta_{zSi} = (\ln(\lambda_{zSi}), \alpha_{zSi})^T$ might plausibly conform to the multivariate normal distribution. The distribution of the difference in median survival in groups with $Z = 1$ and $Z = 2$, $\delta_0 = [\ln(2)/\lambda_{1T0}]^{\alpha_{1T0}} - [\ln(2)/\lambda_{2T0}]^{\alpha^{2T0}}$, can be estimated by simulations from the conditional distribution of θ_{T0} given $\widehat{\theta}_{S0}$ and $\widehat{\psi}$, with bootstrap methods used to account for variability in $\widehat{\psi}$, as in GPHC. Similar methods can be used for piecewise exponential models, as in GPHC. A subtlety arises if S can censor T or T can censor S and the censoring is informative. Then it may be necessary to posit a joint distribution for (T, S), rather than work simply with the marginal distributions, in order to account for informative censoring.

The marginal-level approach can be used for many other types of endpoints (GPHC).

The trial-level correlation R^2_{trial} in equation (9.3) does not depend on within individual correlations, namely correlations between T and S calculable from Σ_{11i} and Σ_{22i}. It is not surprising, therefore, that marginal models yield almost identical estimates of R^2_{trial} as do corresponding bivariate models for T and S (Tibaldi *et al.* 2003, who use the term "univariate" model, instead of marginal model). This is also an indication that marginal models capture most if not all of the surrogate information for predicting treatment effects on T at the trial level. The quantity R^2_{trial} does not account for uncertainty in $\widehat{\psi}$. As pointed out by GPHC, a more realistic measure would be 1 minus the ratio of the variance of $\widehat{\delta}_0$ based on $\widehat{\theta}_{S0}$ and $\widehat{\psi}$, with bootstrap calculations to account for uncertainty in $\widehat{\psi}$, to the variance of $\widehat{\delta}_0$ based only on $\widehat{\mu}_{1T}$ and $\widehat{\mu}_{2T}$, again with bootstrap calculations to account for variability in $\widehat{\mu}_{1T}$ and $\widehat{\mu}_{2T}$. Typically, this assessment of the value of the surrogate will be less optimistic than that provided by R^2_{trial}.

9.4 Discussion

The meta-analytic approach provides an empirical alternative to having to make the strong assumption that T is independent of Z and X given S in order to estimate effects of a new intervention on T from its effects on S. Marginal models that allow one to estimate features of the marginal distributions of T and S in treated and control groups capture most of the available surrogate information on trial-level effects on T, without the need for elaborate bivariate models. Bivariate models may be needed in the presence of informative censoring, however. The ability of the surrogate to predict intervention effects in a new study depends primarily on how tightly summary parameters of the marginal distribution of T are related to such summary parameters for S in a series of studies of interventions similar to the new intervention.

There is a serious price to be paid in loss of precision from the meta-analytic approach. Even with a large number of previous trials to estimate superpopulation parameters and with a large new experiment on the surrogate, the precision of the estimated treatment effect on T in the new study will typically be much less than from a new study with measurements on T itself. This loss of precision is inherent in the irreducible between-study variation, characterized by ϕ. The loss of precision is compounded when there are 10 or fewer previous studies, because an imprecise estimate of the parameters degrades the precision of estimated treatment effects on T considerably.

Apart from precision, several other limitations of the meta-analytic approach should be mentioned (see GPHC):

1. there may be disagreement as to which studies are similar enough to be used in the meta-analysis;
2. published data may not include estimates of Σ_{11i} and Σ_{22i}, requiring the use of unverified assumptions to estimate ϕ;
3. the normal superpopulation model may not be applicable, even after transformation of the parameters θ, and more complex methods may be required for non-normal superpopulations models;
4. stopping the new study early on the basis of surrogate information may restrict the ability of the study to detect unanticipated toxicities of the new treatment; and
5. comprehensive evaluation of a new treatment may require examining several clinical endpoints, so that T becomes a vector. In this case, the use of surrogates becomes more complex and less appealing.

Further methodological research and experience with the method will be needed to determine the extent to which meta-analysis can assist in the evaluation and use of surrogate endpoints.